This is the pre-peer reviewed version of the following article which has been published in final form at: **https://doi.org/10.1002/sim.3998**

Metcalfe C. The analysis of cross-over trials with baseline measurements. Statistics in Medicine 2010; 29: 3211-3218.

The analysis of cross-over trials with baseline measurements

Chris Metcalfe

Department of Social Medicine, Bristol University, Canynge Hall, 39 Whatley Road, Bristol, UK, BS8 2PS

E: chris.metcalfe@bristol.ac.uk

T: +44 (0) 117 928 7326

F: +44 (0) 117 928 7325

Key words: Analysis of covariance (ANCOVA); Baseline observations; Change scores; Cross-over trial; Simulation study

SUMMARY

The statistical power of cross-over trials can be increased by taking “baseline” measurements of the outcome variable at the start of each treatment period. Analysis of covariance (ANCOVA), rather than analysis of change scores, takes best advantage of this. However, ANCOVA can give biased treatment effect estimates in observational studies with true baseline imbalance. Whilst in truth balanced, chance baseline imbalance is possible in individual randomised cross-over studies due to their typically small sample size. Although such chance imbalance does not cause biased estimation on average over repeated trials, this simulation study will aim to confirm the appropriateness of ANCOVA when faced with the analysis of data from an individual trial in which chance baseline imbalance is clearly apparent. Randomised cross-over trials were simulated, varying in sample size and the pattern and strength of correlation between repeated measures. Estimates from ANCOVA, change scores, and post-treatment difference were unbiased on average across each set of simulated datasets. ANCOVA and change scores could use baseline information to improve precision, but change scores could also reduce precision if baseline measures were uninformative. Change scores only were correlated with chance within-subject baseline imbalance. All three estimators could be correlated with chance between-subjects imbalance in the first period baseline measurements, the strongest associations being with the post-treatment difference. Consistent results were obtained from a real data example. In conclusion, ANCOVA took best advantage of baseline measures to improve precision, and avoided bias in the widest set of circumstances with chance imbalance in those baseline measures.

INTRODUCTION

Cross-over trials allow efficient within-subject comparisons of interventions aimed at symptom control, so long as the underlying disease severity remains stable over the successive periods of treatment and separation of treatment periods is sufficient to avoid "carry-over" of an intervention's effects into the following treatment period [1]. In the simplest case of two interventions T and C, and two treatment periods, each trial participant receives each treatment in turn in a randomly allocated order. With a continuous outcome measured at the end of each treatment period, the most straight-forward estimate of the treatment effect is obtained as the mean of the within-subject differences, and a p-value obtained using the paired t-test.

Baseline measurements of the outcome measure at the start of each treatment period can increase the statistical power of a cross-over trial [1,2]. Analysis of covariance (ANCOVA) has long been recommended as taking best advantage of these baseline measures, as this method explicitly estimates the association between baseline and post-treatment measures [1-4]. The alternative, to compare change from baseline between the two treatments, implicitly assumes perfect correlation between baseline and post-treatment measures, and an estimate with poorer precision can result when the correlation is in fact quite low [1,3-5].

ANCOVA may give biased treatment effect estimates in observational studies, where the observed treatment groups differ in case-mix and hence there is a genuine imbalance in the average baseline measures [6]. Sample sizes are typically small in cross-over studies and, despite random allocation to treatment orders, noticeable between-patient imbalance will often be apparent when (for example) first period baseline measures are

compared between the two treatments [7,8]. In addition, within-patient differences in baseline measures may occur through the play of random measurement error or random variation in an individual's symptom levels over time. The aim here is to investigate whether the random allocation of treatments to periods prevents bias in ANCOVA estimates even when baseline imbalance is clearly apparent from the data observed in a particular trial. In other words, is it sufficient for baselines to be balanced in the long run when baselines in a particular trial are obviously not balanced? This question has been addressed previously of course, in the context of both parallel group and cross-over trials [9-11]. The present paper brings a selection of this material together and, using a simulation study, adds further justification to the recommendation for the use of ANCOVA rather than the comparison of change scores.

The simulation study will focus on the analysis of individual cross-over trials, rather than an algebraic analysis of the long-run properties of the estimates. Factors varied in the simulation study will include the sample size, and the inter-correlations between the repeated measures of the outcome. The factor levels examined will be in line with cross-over studies of interventions for asthma and cystic fibrosis with lung function as the outcome measure. The ANCOVA and change from baseline methods will be applied to the simulated data sets, and the mean effect estimates, standard errors, and correlation between baseline imbalance and effect estimates will be examined. The performance of the two statistical methods will also be investigated in an analysis of real data from a cross-over study of treatments for cystic fibrosis.

METHODS

Notation and data set simulation

Baseline (X_{Ti}, X_{Ci}) and post-treatment (Y_{Ti}, Y_{Ci}) values were simulated for n individuals i (i=1, …, n) in a cross-over study comparing active treatment T with control C. For simplicity, it is assumed that systematic bias due to period effects, carry-over, and treatment by period interaction has been avoided. The effect of treatment T is represented by β_T, and the placebo effect of control C as β_C. For all simulations the treatment effect β_T is assumed to be an increase in lung function of 0.1 on the natural log scale, and the placebo effect β_C is assumed to be zero.

Factors varied in different simulations were study size (n = 24, 48), and the pattern and strength of correlation between repeated measures. The pattern of correlations was either an equal correlation between every pair of measurements (an exchangeable correlation matrix) or a diminishing correlation with an increasing separation in time between a pair of measurements (an autoregressive correlation matrix; see the Appendix). The strength of correlation was 0.6 or 0.8 between all pairs of measurements in simulations employing the exchangeable correlation matrix, and between successive measurements in simulations employing the autoregressive correlation matrix before reducing according to a power function (see Appendix). In an additional simulation study active treatment reduced correlations with the post-treatment measure to 0.6, this being achieved by employing what was otherwise an exchangeable correlation matrix with correlation of 0.8 between pairs of measurements. A further simulation based on the autoregressive correlation matrix had the correlation between successive measurements reduced by the washout period (see Appendix).

Ten thousand datasets were simulated for each combination of factor levels considered. The four observed values were simulated for each patient as realisations from correlated normally distributed variables with, marginally, mean and standard deviation 0.4 using the functions presented in the Appendix. All these values are realistic for cross-over studies of interventions aimed at improving lung function [5,12,13].

Statistical methods and measures of performance

The ANCOVA estimate is the intercept term in a normal errors regression of (Y_{Ti} - Y_{Ci}) on (X_{Ti} - X_{Ci}), this approach being exactly equivalent in this simple context to that proposed elsewhere [1,5]. The change scores estimate is the mean of the differences between (Y_{Ti} - X_{Ti}) and (Y_{Ci} - X_{Ci}). The post-treatment only estimate is the mean of differences between Y_{Ti} and Y_{Ci}.

The performance of each statistical measure was measured as the average effect estimate, and its standard error estimated empirically as the standard deviation over 10,000 simulations. In addition, the correlations between within-subject baseline imbalance, between-subject period 1 baseline imbalance, and each treatment effect estimate over each set of 10,000 simulations were calculated.

Real data example

Data was made available from a three period, three treatment cross-over study of daily rhDNase, alternate day rhDNase and hypertonic saline, the aim being to improve lung function in children with cystic fibrosis [12,13]. Treatment periods of twelve weeks

were separated by two week washout periods. Data on the primary outcome measure, forced expiratory volume in one second (FEV_1) was used in the present analysis. The two pair-wise comparisons made in the original analysis are repeated here using the statistical methods under consideration: hypertonic saline versus daily rhDNase and alternate day rhDNase versus daily rhDNase. Log-transformed measures of FEV_1 were used to allow percentage changes in lung function to be estimated, and that approach is adopted in the present investigation.

RESULTS

Table 1 presents the results of the simulation study employing an exchangeable correlation matrix. The means over the 10,000 simulations show each of the three methods to give unbiased estimates in the long run. The standard errors indicate that the poorest precision is achieved by the change scores, but that there is no clear difference between the ANCOVA and post-treatment only estimates in these simulation studies. Estimates from the three different measures are inter-correlated, with a correlation in each of the different simulation scenarios of around 0.9 between ANCOVA and the post-treatment difference, around 0.7 between ANCOVA and change scores, and around 0.7 between change scores and the post-treatment difference. The chance occurrence of within-subject baseline imbalance between treatments is negatively correlated with the change scores estimates, but not with the ANCOVA and post-treatment only estimates (Table 1). The association is more apparent in the smaller studies, and appears to be unrelated to the strength of association between repeated measurements. A similar pattern of correlations is apparent between treatment effect

estimates and between-subject baseline imbalance in the first period (Table 1), except that the correlation between this imbalance and the change scores was greater when the correlations between repeated measures were weakened. Very similar results are obtained when active treatment reduces the correlation between the post-treatment measurement and other measurements.

Table 2 presents the results of the simulation studies based upon an autoregressive correlation matrix. Looking first at the simulations without an extra effect of the washout period (first three rows of Table 2) there is no convincing evidence of biased estimation, and again the poorest precision in estimation is achieved by change scores, with little to choose between ANCOVA and post-treatment only estimates. The correlation between the ANCOVA and change score estimates is around 0.8 in these simulations, around 0.9 between ANCOVA and post-treatment only estimates, and around 0.55 between change scores and post-treatment only estimates. Chance within-subject baseline imbalance between the two treatments is negatively correlated with change score estimates, positively correlated with post-treatment only estimates, but is not clearly correlated with ANCOVA estimates in these simulations. The strength of correlation with post-treatment only estimates reduces with both increasing sample size and with weaker correlations between repeated measures, whilst the strength of correlation with change scores reduces with increasing sample size only. In these simulations all three sets of estimates correlate with between-subject baseline imbalance at the start of the first period, with ANCOVA and post-treatment only estimates being positively associated with the imbalance, and change scores being negatively associated (Table 2, Figure 1). These correlations are not strongly influenced by the sample size, or by the strength of correlation between repeated measures.

Results in the lowest two rows of Table 2 arise from simulations based on autoregressive correlations between repeated measures, but with the washout period causing the correlations between the two treatment periods to be weaker. The average observed correlation between two measurements in the same treatment period was 0.80, whereas that between pairs of measurement in different treatment periods was 0.51. In these simulations use of the baseline measures does improve the precision of treatment effect estimates, with the ANCOVA estimates offering a small advantage over change scores. Correlations of the treatment effect estimates with within-subject and between-subject differences in baseline values were unaffected by the weakening of between-period correlations during the washout period.

Table 3 presents the results of the real data example. The expected greater precision of the ANCOVA compared to the change score estimates is observed but there was no improvement in precision with the use of baseline measures. For the comparison of daily and alternate day rhDNase, there was little within-subject difference between the pre-treatment measures, and as was consequently expected no important differences in the estimates of the treatment effect. There was a larger within-subject difference in pre-treatment measures for the comparison between rhDNase and hypertonic saline and greater variation in the estimates resulting from the different statistical methods. The change score estimate was most discrepant from the ANCOVA estimate, suggesting a larger advantage of daily rhDNase compared to hypertonic saline. Repeated measurements of FEV_1 were very highly correlated, being around 0.9 for successive measures to around 0.85 for measurements taken some time apart. Table 4 gives the full correlation matrix which could be considered to fall part way between an exchangeable and an autoregressive pattern.

DISCUSSION

In the long-run all three statistical methods provided unbiased estimates of the true treatment effect in the simulation studies considered. The expected extra efficiency with the use of baseline measures was only clearly observed in simulations where the washout period reduced the strength of correlation between measures taken in different treatment periods, with ANCOVA appearing best able to capitalise on that extra efficiency. Across the range of simulations in the present investigation, ANCOVA was found least likely affected by chance baseline imbalance between treatments, a marked correlation with between-subject imbalance in first period baselines only being apparent in simulations based on autoregressive correlation between repeated measurements. Even in that situation, the simple difference in post-treatment measurements was more strongly correlated with that between-subject imbalance. In the real data example improved efficiency with the use of baseline measures was not apparent, with their use in change scores impacting negatively to give poorer precision. The effect of baseline imbalance was consistent with the simulation study results, with noticeable differences in the resulting treatment effect estimates only for the comparison with a moderate within-subject imbalance in the baseline measures.

An improvement in treatment effect estimate precision was not apparent with the use of baseline measurements in the real data example, and only in those simulations where the washout period weakened the correlations between measurements taken in different treatment periods. Seeing “in the flesh” the limited the circumstances in which collecting extra information improved precision does cause surprise, even though this

pattern of results was predictable. Only when the correlation between measurements taken in the same treatment period exceeds that between a pair of measurements taken in different treatment periods do baseline measurements bring extra precision to a cross-over study [1,2]. This is not the case for simulations generated using an exchangeable correlation matrix, and is only clearly apparent in the simulations generated using the autoregressive correlation matrix once the washout period reduces the correlation between successive measurements. Hence baseline measurements may be most useful in randomised controlled trials requiring a long washout period, in these circumstances contributing to the reduction in error variation which the within-subject difference in distant post-treatment measurements may be less able to do. Such an effect was not clearly apparent in the real data example, where twelve week long treatment periods were separated by relatively short two week long washout periods.

The differing results with the different methods applied to the comparison of daily rhDNase and hypertonic saline in the real data example can be understood from the well-known relationships between the methods [3,6,11]. Considering our implementation of the ANCOVA method:

$$(Y_{Ti} - Y_{Ci}) = \beta_T + \gamma(X_{Ti} - X_{Ci}) \qquad (1)$$

So in the comparison of daily rhDNase and hypertonic saline $(X_{Ti} - X_{Ci})$ was negative as the hypertonic saline group had better average lung function at baseline. That fact is ignored by the post-treatment only analysis, with γ set to zero, and with highly correlated measurements the estimated advantage of daily rhDNase β_T was attenuated by the perseverance of the chance baseline difference into the post-treatment period. With the comparison of change scores, γ is set to 1 and β_T was estimated as the

difference in post-treatment measures, inflated to compensate for the whole disadvantage for treatment *T* at baseline. For ANCOVA γ is estimated, in practical circumstances as a value between 0 and 1 reflecting the strength of association between baseline and post-treatment measures, and β_T was inflated to account for that part of the baseline disadvantage for treatment *T* which carried over to post-treatment. So with no allowance for the baseline disadvantage for daily rhDNase, the post-treatment only estimate of β_T was the most modest, with partial allowance for that disadvantage the ANCOVA estimate was greater, and with complete, perhaps over-compensation for the disadvantage, the change scores estimate was the highest by some way.

With cross-over studies having relatively modest sample sizes, random allocation still allows a high chance of noticeable between-subject imbalance in the first baseline measurement, and within-subject imbalance between the pairs of baseline measurements. Within-subject imbalance in baseline measures is captured as part of the change scores and ANCOVA approaches, but the change scores method is based on the whole of the baseline difference feeding through to the post-treatment measurements (equation 1 above), and so will usually over-correct and give biased treatment effect estimates. Between-subject imbalance in the first period baseline measurement is not directly measured by any of the three statistical methods considered here. Consequently, in the present simulations based on autoregressive correlations between repeated measures, the post-treatment difference demonstrates the full effect of this baseline imbalance, which is under-corrected by the ANCOVA approach and over-corrected by the analysis of change scores (see Figure 1). Overall, the ANCOVA method gave biased estimates in the most restricted set of circumstances, so providing further justification to the use of this method. Change scores were correlated with within-subject and between-

subject imbalance in all circumstances covered by the simulation study, whilst the post-treatment difference showed a similar pattern to ANCOVA but with a stronger association with between-subject imbalance when repeated measures were generated using an autoregressive correlation structure.

In conclusion, the long-standing recommendation for the use of ANCOVA, rather than change scores, in the analysis of cross-over studies with baseline measurements is supported by the findings of the present study. The method will take best advantage of any further information in baseline measurements to improve precision of estimation, but will not reduce that precision when the baselines are uninformative. Furthermore the present study shows ANCOVA as being the most likely method to avoid bias when there is chance within-subject or between-subject imbalance in the baseline measures. The practical consequence of these findings is that the ANCOVA method can safely be pre-specified in the statistical protocols of planned cross-over studies, as it was for the example study considered here.

ACKNOWLEDGEMENTS

I am grateful to Ranjan Suri (Great Ormond Street Hospital for Children, London), Colin Wallis (Great Ormond Street Hospital for Children, London) and Andrew Bush (Royal Brompton Hospital, London) for permission to use the data from the study of daily rhDNase, alternate day rhDNase and hypertonic saline in the treatment of cystic fibrosis. That study was funded by the UK NHS Health Technology Assessment programme. Many thanks to Lisa Hampson (Bristol University) for useful comments on the original draft manuscript; any errors in the current version are solely my responsibility. In addition, my attempts to incorporate the suggestions of an anonymous referee greatly improved the manuscript. The results of an earlier version of the simulation study were the basis of a poster presentation at the International Society for Clinical Biostatistics (ISCB) conference in Prague, 2009.

REFERENCES

1. Senn S. *Cross-over trials in clinical research. 2nd Edition.* Wiley: Chichester, 2002;62-67.

2. Kenward MG, Roger JH. The use of baseline covariates in crossover studies. *Biostatistics* 2010; **11**: 1-17. DOI: 10.1093/biostatistics/kxp046

3. Hills M, Armitage P. The two-period cross-over clinical trial. *British Journal of Clinical Pharmacology* 1979; **8**: 7-20

4. Fleiss JL. A critique of recent research on the two-treatment crossover design. *Controlled Clinical Trials* 1989; **10**: 237-243

5. Kenward MG, Jones B. The analysis of data from 2 x 2 cross-over trials with baseline measurements. *Statistics in Medicine* 1987; **6**: 911-926

6. van Breukelen GJ. ANCOVA versus change from baseline: more power in randomized studies, more bias in nonrandomized studies. *Journal of Clinical Epidemiology* 2006; **59**: 920-925. DOI: 10.1016/j.jclinepi.2006.02.007

7. Grender JM, Johnson WD, Elston RC. Regression toward the mean in 2 x 2 crossover designs with baseline measurements. *Statistics in Medicine* 1992; **11**: 727-741

8. Grender JM, Johnson WD, Elston RC. Authors' reply. *Statistics in Medicine* 1993; **12**: 1088-1089

9. Senn SJ. Covariate imbalance and random allocation in clinical trials. *Statistics in Medicine* 1989; **8**: 467-475

10. Senn S. In defence of analysis of covariance: a reply to Chambless and Roeback. *Statistics in Medicine* 1995; **14**: 2283-2285

11. Senn S. Change from baseline and analysis of covariance revisited. *Statistics in Medicine* 2006; **25**: 4334-4344. DOI: 10.1002/sim.2682

12. Suri R, Wallis C, Bush A, Thompson S, Normand C, Flather M, Grieve R, Metcalfe C, Lees B. A comparative study of hypertonic saline, daily and alternate-day rhDNase in children with cystic fibrosis. *Health Technology Assessment* 2002; **6**: 1-60

13. Suri R, Metcalfe C, Lees B, Grieve R, Flather M, Normand C, Thompson S, Bush A, Wallis C. Comparison of hypertonic saline and alternate-day or daily recombinant human deoxyribonuclease in children with cystic fibrosis: a randomised trial. *Lancet* 2001; **358**: 1316-1321. DOI: 10.1016/S0140-6736(01)06412-1

APPENDIX: Simulation methods

Exchangeable correlation matrix

These simulations are based on an exchangeable correlation matrix to describe the relationships between the four repeated measures:

$$\begin{pmatrix} 1 & & & \\ \rho^2 & 1 & & \\ \rho^2 & \rho^2 & 1 & \\ \rho^2 & \rho^2 & \rho^2 & 1 \end{pmatrix}$$

where successive rows and columns correspond to the successive measurements. Box A1 gives the functions used in the simulation of values for an individual *i*, where $rand_0$, $rand_1$, $rand_2$, $rand_3$ and $rand_4$ are five randomly generated values from a standard normal variable. *X* are simulated baseline measurements, *Y* are simulated outcome measurements, *C* denotes the control group and *T* the treatment group. Here ρ is the square root of the resulting correlation between pairs of simulated measures, values of 0.894427 (square root of 0.8) and 0.7746 (square root of 0.6) being employed in this study. The true mean and standard deviation of the simulated measurements are denoted as μ and σ respectively, 0.4 being used for both in all of these simulations. The true treatment effect is denoted as *trt* and is 0.1 in all simulations. In some simulations the treatment reduces the correlation between baseline and post-treatment measurements, with π changed from 1 to 0.8 to introduce this aspect.

Autoregressive correlation matrix

These simulations are based on a first order autoregressive correlation matrix to describe the relationships between repeated measures:

$$\begin{pmatrix} 1 & & & \\ \rho & 1 & & \\ \rho^2 & \rho & 1 & \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$$

where successive rows and columns correspond to the successive measurements. Box A2 describes the functions used in the simulation of values for an individual *i*, where $rand_1$, $rand_2$, $rand_3$ and $rand_4$ are four randomly generated values from a standard normal variable; and m_1, m_2, m_3, and m_4 are intermediate calculations which introduce the autoregressive correlation across the sequence of measurements. *X* are simulated baseline measurements, *Y* are simulated outcome measurements, *C* denotes the control group and *T* the treatment group. Here ρ is the correlation between one measure and that immediately following it, the correlation between pairs of measurements being less if they are further apart in the sequence of measurements. Taking this a step further, additional simulations were conducted with the washout period weakening the correlation between m_2 and m_3 further, according to the correlation matrix:

$$\begin{pmatrix} 1 & & & \\ \rho & 1 & & \\ \rho^3 & \rho^2 & 1 & \\ \rho^4 & \rho^3 & \rho & 1 \end{pmatrix}$$

The true mean and standard deviation are denoted as μ and σ respectively with both taking the value of 0.4 in all simulations. The true treatment effect is denoted *trt* and is assumed to be 0.1 in all simulations.

Figure 1. Effect estimates from the three methods plotted against the between-subject difference in period one baseline measures for 10,000 simulated datasets generated according to the "weaker correlation" scenario in Table 2, with the regression line (dashed)

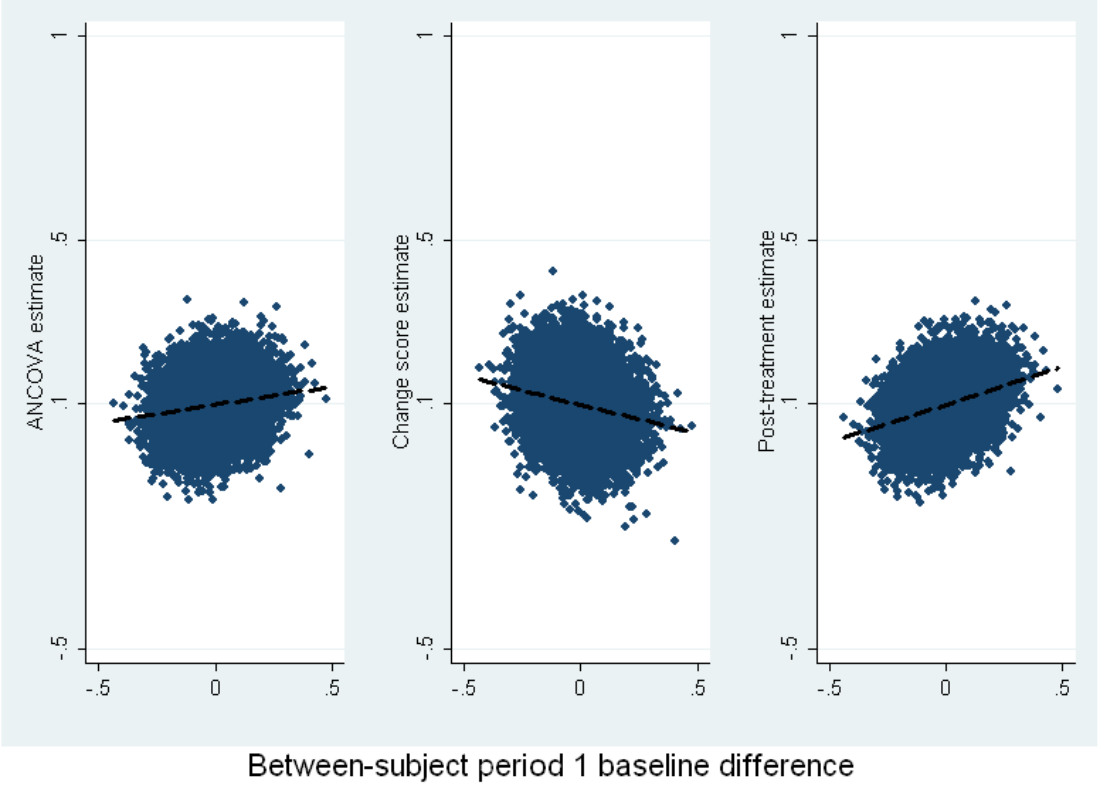

Table 1. Results of the simulation study, based on an exchangeable correlation matrix. For each combination of factor levels 10,000 data sets were simulated with a true treatment effect of 0.1

	Effect estimate	(Standard error)	Correlation with within-subject baseline difference	Correlation with between-subject period 1 baseline difference	Factor levels
ANCOVA	0.100	(0.052)	-0.007	+0.005	**MODERATE STUDY**
Change scores	0.101	(0.072)	-0.150	-0.221	n=24, correlation=0.8,
Post-treatment only	0.100	(0.051)	-0.008	+0.004	SD(within-subject baseline diff.) = 0.255
ANCOVA	0.100	(0.037)	-0.008	-0.002	**LARGER STUDY**
Change scores	0.099	(0.051)	-0.098	-0.228	n=48, correlation=0.8,
Post-treatment only	0.100	(0.036)	-0.008	-0.004	SD(within-subject baseline diff.) = 0.250
ANCOVA	0.100	(0.052)	-0.008	-0.007	**WEAKER CORRELATION**
Change scores	0.100	(0.073)	-0.098	-0.322	n=48, correlation=0.6,
Post-treatment only	0.100	(0.051)	-0.008	-0.010	SD(within-subject baseline diff.) = 0.353
ANCOVA	0.100	(0.049)	-0.009	-0.008	**TREATMENT REDUCES CORRELATION**
Change scores	0.099	(0.061)	-0.085	-0.196	n=48, correlation=0.8, reduced to 0.64,
Post-treatment only	0.100	(0.049)	-0.010	-0.010	SD(within-subject baseline diff.) = 0.250

Table 2. Results of the simulation study, based on an autoregressive correlation matrix. For each combination of factor levels 10,000 data sets were simulated with a true treatment effect of 0.1

	Effect estimate	(Standard error)	Correlation with within-subject baseline difference	Correlation with between-subject period 1 baseline difference	Factor levels
ANCOVA	0.100	(0.064)	-0.004	+0.167	**MODERATE SIZE STUDY**
Change scores	0.100	(0.075)	-0.123	-0.089	n=24, correlation=0.8^j,
Post-treatment only	0.100	(0.068)	+0.083	+0.331	SD(within-subject baseline diff.) = 0.345
ANCOVA	0.099	(0.045)	-0.016	+0.168	**LARGER SIZE STUDY**
Change scores	0.098	(0.053)	-0.100	-0.095	n=48, correlation=0.8^j,
Post-treatment only	0.099	(0.049)	+0.047	+0.328	SD(within-subject baseline diff.) = 0.340
ANCOVA	0.098	(0.063)	-0.017	+0.160	**WEAKER CORRELATION**
Change scores	0.098	(0.077)	-0.103	-0.206	n=48, correlation=0.6^j,
Post-treatment only	0.098	(0.065)	+0.028	+0.324	SD(within-subject baseline diff.) = 0.453
ANCOVA	0.100	(0.067)	-0.005	+0.125	**WASHOUT WEAKENS CORRELATION**
Change scores	0.100	(0.075)	-0.106	-0.116	n=24, correlation=0.8^j,
Post-treatment only	0.100	(0.080)	+0.117	+0.388	SD(within-subject baseline diff.) = 0.401
ANCOVA	0.099	(0.047)	-0.015	+0.124	**WASHOUT WEAKENS CORRELATION**
Change scores	0.098	(0.053)	-0.084	-0.123	n=48, correlation=0.8^j,
Post-treatment only	0.099	(0.057)	+0.071	+0.385	SD(within-subject baseline diff.) = 0.395

Table 3. Results from the real data example. Positive effect estimates indicate a higher or more improved lung function with daily rhDNase

	Daily versus alternate-day rhDNase			Daily rhDNase versus hypertonic saline		
	Effect estimate	(Standard error)	p-value	Effect estimate	(Standard error)	p-value
Pre-treatment (within-patient)	*0.0171*	*(0.0296)*	*0.566*	*-0.0688*	*(0.0334)*	*0.046*
ANCOVA	0.0199	(0.0326)	0.546	0.0744	(0.0291)	0.015
Change scores	0.0047	(0.0415)	0.911	0.1283	(0.0378)	0.002
Post-treatment only	0.0218	(0.0323)	0.503	0.0595	(0.0282)	0.041

Table 4. Correlation matrix of observed associations between repeated measures (log scale) in the real data example, a three period cross-over study.

Measurement	1	2	3	4	5	6
1	1					
2	0.92	1				
3	0.94	0.93	1			
4	0.85	0.89	0.83	1		
5	0.88	0.89	0.85	0.84	1	
6	0.84	0.86	0.78	0.90	0.84	1

Box A1. Simulations based upon an exchangeable correlation matrix

Simulated measure	Function used to simulate the measure from the randomly generated values
X_{Ci}	$((\rho \times rand_0 + (1 - \rho^2)^{0.5} \times rand_1) \times \sigma) + \mu$
Y_{Ci}	$((\rho \times rand_0 + (1 - \rho^2)^{0.5} \times rand_2) \times \sigma) + \mu$
X_{Ti},	$((\rho \times rand_0 + (1 - \rho^2)^{0.5} \times rand_3) \times \sigma) + \mu$
Y_{Ti}	$((\pi\rho \times rand_0 + (1 - \pi\rho^2)^{0.5} \times rand_4) \times \sigma) + \mu + trt$

Box A2. Simulations based upon an autoregressive correlation matrix

Simulated measure	Functions used to simulate the measure from the randomly generated values
m_1	$rand_1$
m_2	$(\rho \times m_1 + (1 - \rho^2)^{0.5} \times rand_2)$
m_3	$(\rho \times m_2 + (1 - \rho^2)^{0.5} \times rand_3)$
m_4	$(\rho \times m_3 + (1 - \rho^2)^{0.5} \times rand_4)$
Allocated to control followed by treatment:	
X_{Ci}	$(m_1 \times \sigma) + \mu$
Y_{Ci}	$(m_2 \times \sigma) + \mu$
X_{Ti},	$(m_3 \times \sigma) + \mu$
Y_{Ti}	$(m_4 \times \sigma) + \mu + trt$
Allocated to treatment followed by control:	
X_{Ci}	$(m_3 \times \sigma) + \mu$
Y_{Ci}	$(m_4 \times \sigma) + \mu$
X_{Ti},	$(m_1 \times \sigma) + \mu$
Y_{Ti}	$(m_2 \times \sigma) + \mu + trt$